

MIS536X, Fall 2014
Assignment 0

Shakespeare Analytics
(Due: 5pm, 09/18)

1. Introduction

The purpose of this simple assignment is to help students brush up their Java programming skill and lead into course material.

Please refer to the course syllabus for assignment-related policies. This is an individual assignment, therefore any collaboration or plagiarism will be heavily penalized.

2. Algorithm and Dataset

Attached is a subset of Shakespeare's work, in plain text format. You are required to write a simple Java program that identifies the top 50 frequent trigrams (contiguous sequence of 3 words). Please ignore all XML tags, numbers, and punctuation marks.

Requirements:

- Make (and state) reasonable assumptions about text cases
- The program should have reasonably professional style and comments
- The program should be executable in a command line environment (i.e., independent of a particular IDE)
- No hard coding of file names in the program

3. What to Turn in

You should turn in the following, zipped in a single file for submission on Blackboard:

- The Java source code
- A CSV file that lists the top 50 trigrams (in descending order of frequency), accompanied by actual frequency counts.

4. Grading

You will be evaluated on:

- Correctness of your code
- Elegance of your code
- Correctness of the output